By: Abraham Pabbati

¿Por qué Lakehouse sobre el almacén de datos?

Introducción

Muchos de ustedes que han visto mi perfil de LinkedIn deben estar preguntándose qué significa el título de mi perfil. Durante mucho tiempo, un almacén de datos ha sido el estándar de oro para integrar datos de sistemas dispares en una sola base de datos para analizar y obtener información comercial. Después de haber trabajado en el espacio de análisis de datos durante más de 20 años, nunca he estado más entusiasmado con una nueva tecnología/arquitectura que realmente creo que es un avance material en la forma en que administramos y aprovechamos los datos, es decir, Lakehouse .

Para aquellos de ustedes que no están familiarizados con Lakehouse, aquí hay una definición simplificada. Lakehouse es una combinación de las mejores características de unLago de datosy unAlmacén de datos. Un lago de datos proporciona almacenamiento de bajo costo altamente escalable paraestructurado y no estructuradodata, mientras que un Data Warehouse proporciona consultas de alto rendimiento sobre datos estructurados. Lakehouse ofrece las dos características anteriores.

A continuación, proporciono cuatro razones por las que creo que Lakehouse es lo que debemos construir en el futuro en lugar de un almacén de datos.

1. Menos saltos: cargas de datos optimizadas/más rápidas

La mayoría de las empresas tienen actualmente tres tipos de sistemas de almacenamiento de datos.

a) Bases de datos de aplicaciones: sistemas transaccionales que capturan datos de todas las operaciones de la empresa, por ejemplo, recursos humanos, finanzas, CRM, ventas, etc.

b) Lagos de datos: estos son sistemas de almacenamiento en la nube comodín que almacenan datos estructurados y no estructurados, como copias de seguridad de datos de aplicaciones, registros, secuencias de clics web, imágenes, videos, etc.

c) Almacenes de datos: datos limpios e integrados organizados de manera que mejoren el rendimiento de las consultas para que podamos ejecutar informes y paneles rápidamente.

La mayoría de los equipos de ingeniería de datos, cuando construyen canalizaciones de datos, normalmente mueven los datos de las bases de datos de aplicaciones a los lagos de datos y luego mueven un subconjunto de los datos a un almacén de datos para fines de generación de informes.

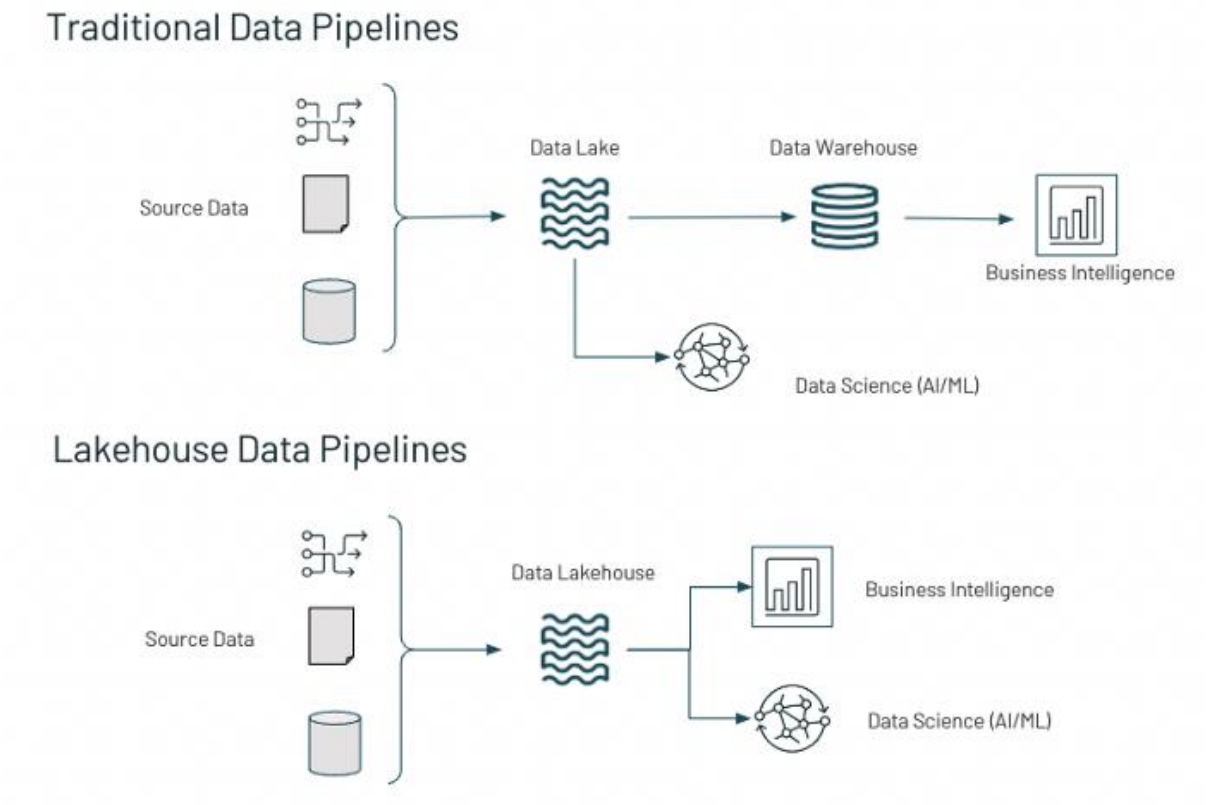

Fig. 1: Canalizaciones de datos tradicionales frente a canalizaciones de Data Lakehouse

Pero con la arquitectura Lakehouse, combinamos Data Lake y Data Warehouse en Lakehouse, de modo que sus datos se muevan a través de solo dos tipos de sistemas. Además, Lakehouse puede admitir la ingesta de datos de Streaming y Batch en la misma estructura de datos. Esto evita el paso adicional para consolidar datos por lotes y de transmisión. Esto da como resultado una canalización de datos más optimizada con la menor cantidad de saltos y un tiempo de obtención de valor más rápido. Por lo tanto, con Lakehouse, en la mayoría de los casos, lleva menos de 5 minutos obtener sus datos desde el punto en que se generaron hasta el punto en que se pueden informar en un tablero.

2. Pasado y futuro: única fuente de verdad

A medida que crece la madurez de los equipos de datos, las empresas no están satisfechas con solo crear informes de inteligencia comercial (BI) tradicionales, sino que muchas empresas ahora tienen científicos de datos internos que usan inteligencia artificial y algoritmos de aprendizaje automático (AI/ML) para descubrir patrones ocultos dentro de sus datos para predecir el futuro. La mayoría de estos científicos de datos confían en Python, un lenguaje de programación, para construir sus modelos AI/ML y requieren grandes cantidades de datos de entrada para construir modelos confiables. Esto ha llevado al crecimiento de los lagos de datos que normalmente contienen los datos que estos científicos de datos necesitan para construir sus modelos (como se muestra en la figura 1). Pero cuando los datos de una empresa se distribuyen en lagos de datos y almacenes de datos, los científicos de datos tienen que buscar sus datos en dos sistemas diferentes y reunirlos en un lago de datos y luego comenzar el proceso de creación de sus modelos. Además, los almacenes de datos normalmente no admiten el lenguaje python y solo admiten la

interfaz SQL , lo que dificulta que los científicos de datos los aprovechen como fuente de datos.

Al integrar Data Warehouses y Data Lakes en un solo Lakehouse que admite tanto SQL como Python, los científicos de datos ahora pueden consultar un solo sistema (es decir, Lakehouse) que ahora actúa como la única fuente de verdad para los datos de toda la empresa. Esto simplifica y acelera el proceso de creación de informes de BI, así como modelos de IA/ML.

3. Código abierto: sin bloqueo de proveedores

Entonces, ¿qué hace que Lakehouse sea superior a un simple Data Lake? Bueno, todo se reduce a la forma en que se almacenan los datos.

a) Formato: en lugar de que los datos se almacenen en formatos simples legibles por humanos, como csv o json, en Lakehouse, almacenamos los datos en un formato de parquet comprimido que es mejor para que las computadoras lean y escriban datos rápidamente. Además, el parquet es un formato de código abierto, por lo que sus datos no están bloqueados en un formato propietario. Puede usar el lenguaje python para acceder a estos datos desde su computadora portátil sin tener que pagarle a un proveedor para acceder a sus propios datos.

b) Registros: además de almacenar los datos en formato parquet, también realizamos un seguimiento de los metadatos de los archivos de parquet y mantenemos un registro de todas las operaciones ejecutadas en estos archivos. Esto nos permite organizar y administrar mejor los datos dentro de estos archivos. Si bien la descripción completa de cómo ayudan estos archivos de registro está más allá del alcance de este artículo, basta con decir que estos archivos de registro/metadatos son el ingrediente secreto (bueno, no tan secretos ya que son de código abierto

) que hace que la experiencia de administrar datos en un Lakehouse mucho más fácil y eficiente.

Actualmente hay tres grandes proyectos de código abierto que ayudan a construir una casa del lago, a saber, delta lake , iceberg y hudi . Cada uno de estos formatos cuenta con el apoyo de las principales empresas, así como de desarrolladores comunitarios individuales.

4. Pago por evento: bajo costo

Finalmente, el cuarto y último punto que quería hacer era sobre el costo de almacenar y consultar los datos. Dado que los datos se almacenan en sistemas de almacenamiento en la nube económicos, como AWS S3, Azure ADLS o GCP GCS , el costo de almacenar un TB de datos durante un mes suele oscilar entre $20 y $50, según sus requisitos de recuperación ante desastres y alta disponibilidad. . El otro costo en el que incurre es el costo de la computación, que generalmente es un Apache Spark Cluster que cuesta alrededor de $ 2 por hora. Para un pequeño racimo. Para cargas de trabajo más pequeñas, ni siquiera necesita un clúster Spark, puede consultar los datos desde su computadora portátil o una máquina virtual en la nube utilizando el lenguaje python. A diferencia de los almacenes de datos, que suelen estar siempre activados, incluso cuando no los está utilizando, en una arquitectura Lakehouse, activa el clúster solo cuando necesita consultar los datos. Si bien los almacenes de datos basados en la nube más nuevos brindan poder de cómputo bajo demanda, los puntos de precio suelen ser más altos que los de Spark Clusters y no le permiten consultar sus datos desde su computadora portátil sin un clúster en funcionamiento.

En conclusión.

Lakehouse lo ayuda a simplificar el procesamiento de datos y democratizar el uso de los datos en toda su organización al menor costo posible. Este es un cambio de juego para las pequeñas y grandes empresas que se están quedando atrás en la búsqueda de aprovechar los datos para obtener una ventaja competitiva. Así que diga adiós a los almacenes de datos y salude a Lakehouse y nunca mirará hacia atrás, ya que será el superhéroe para todos sus usuarios de datos.